IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 December 2022, accepted 12 January 2023, date of publication 16 January 2023, date of current version 24 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3237257

RESEARCH ARTICLE

Resource Allocation With Edge-Cloud Collaborative Traffic Prediction in Integrated Radio and Optical Networks

BOWEN BAO[1], HUI YANG[1], (Senior Member, IEEE), QIUYAN YAO[1], (Member, IEEE), LIN GUAN[1], JIE ZHANG[1], (Member, IEEE), AND MOHAMED CHERIET[2], (Senior Member, IEEE)

[1]State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of System Engineering, École de technologie supérieure (ÉTS), University of Quebec, Montreal, QC H3C 1K3, Canada

Corresponding author: Hui Yang (yanghui@bupt.edu.cn)

This work was supported in part by NSFC Project under Grant 62271075 and Grant 62201088, and in part by the Fund of State Key Laboratory (SKL) of Information Photonics and Optical Communications (IPOC) [Beijing University of Posts and Telecommunications (BUPT)] under Grant IPOC2021ZT04 and Grant IPOC2020A004.

ABSTRACT By integrating communications in different domains, integrated radio and optical networks can serve a wider range of applications and services. Integrated radio and optical network scenarios will involve more weak-computation-ability network nodes, such as small-cell base stations. To pursue efficient integrated radio and optical networks, more efficient ways to conduct transmission under the demand of edge and cloud collaboration are required. The lack of forward-looking resource allocation may easily lead to a waste of network resources without an expected return. Therefore, an efficient resource allocation scheme needs to consider certain issues: 1) a comprehensive perspective of traffic prediction; 2) a release of pressure on the transmission pipeline during the prediction process; and 3) a reduction of loss of edge nodes due to the computation. In this paper, benefiting from machine learning, we propose a resource allocation with edge-cloud collaborative traffic prediction (TP-ECC) in integrated radio and optical networks, where an efficient resource allocation scheme (ERAS) is designed based on the prediction results with the gated recurrent unit model. We maximize the utilization of limited resources to improve the awareness of network status. We present three evaluation indicators and build a network architecture to evaluate our resource allocation scheme. Through edge-cloud collaboration, our proposal can improve traffic prediction accuracy by 9.5% compared with single-point traffic prediction, and resource utilization is also improved by edge-cloud collaborative traffic prediction.

INDEX TERMS Integrated radio and optical networks, resource allocation, edge-cloud collaboration, traffic prediction.

I. INTRODUCTION

The integrated radio and optical networks can serve diversified applications and services by introducing the Internet of Things (IoT) supporting technology which can provide seamless interconnection among heterogeneous devices [1]. With the access of a large number of network devices, a high volume of data would be stored or processed at the edge of weak-computation-ability nodes in integrated radio and optical networks. The architecture of mobile edge computing (MEC) with cloud platform and edge nodes become a new and attractive computing paradigm, which integrates the computing power of the cloud platform with the flexible tasks of edge nodes. It can support various computationally complex delay-sensitive service applications, such as face recognition, natural language processing, and interactive games [2], [3], [4]. Therefore, the MEC architecture has become a typical networking mode for integrated radio and optical networks. However, the available resources in a single edge node (such as a small cell base station) are very limited, which is still an important issue in this scenario [5]. Although there have been some works to upload computing tasks that exceed the

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

capacity of edge node to the remote cloud, the total resource consumption in the system may be high due to the bandwidth occupation on the transmission pipeline [6], [7]. In other words, the limited resource of a single edge node severely degrades the performance of integrated radio and optical networks.

A reasonable resource allocation algorithm can achieve load balancing among edge nodes, which enables the resource-constrained edge nodes to help each other to realize more flexible sharing for workloads and resources [8]. In computing-intensive tasks, it can meet the heterogeneous needs of access terminals [9]. For resource allocation, traffic prediction can be regarded as a key and important first-hand operation that would be regrettable to be skipped. High-precise traffic prediction can instruct people to not only make flexible switch adjustments but also form backup paths in burst traffic [10], which may further enable them to break the limitation in integrated radio and optical networks, so as to improve resource utilization and reduce the blocking rate as well as average queue delay.

Machine learning has been applied to the present traffic prediction, including Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), and Hierarchical Temporal Memory (HTM) [11], [12], [13], which can iteratively predict the next traffic flow of network in time series. Although it has received great attention, most of the existing solutions simplify the assumption that the perception of network state is only based on the commanding perspective of the cloud platform without taking into account those of edge nodes. Motivated by this, we consider and forecast traffic from the perspective of both cloud and edge nodes. Compared with the case where the neural network is executed in a single location, we propose an edge-cloud collaborative traffic prediction (TP-ECC), which is more promising in obtaining higher prediction accuracy. Besides, considering the limited computing capacity in edge nodes, a simple traffic prediction model is required to support the function, which also needs to support the rapid processing of large-scale datasets in cloud nodes. Thus, a gated recurrent unit (GRU) is adopted to forecast the traffic in the TP-ECC model, which has a simpler gate structure than the LSTM traffic prediction model [14]. With the designed edge-cloud collaboration system architecture, the TP-ECC module may improve the accuracy of network state perception, and give an accurate forecast of edge node traffic.

Based on the proposed TP-ECC, we further propose a new efficient resource allocation scheme (ERAS) in integrated radio and optical networks. On the basis of node traffic prediction results, ERAS uses the load balancing theory to allocate the resources among the edge nodes in order to carry end-to-end services. The proposed ERAS algorithm together with the traffic prediction can make continuous strategic adjustments based on the real-time network status, formulate an optimal strategy that meets the needs of users, and send the results to quickly execute network configuration in the device. This optimized resource allocation method provides a new solution for service-oriented network intelligent optimization configuration. We also design a control experiment in the simulation environment to demonstrate the effectiveness of the whole system architecture and strategy through data analysis. The performance evaluation with the simulation of ERAS shows that the accuracy of TP-ECC module performs well. Under its guidance, the performances of ERAS are generally better than that of the previous works.

The main contribution of this paper can be summarized in the following aspects.

1) A system architecture is set up to support edge cloud collaboration in integrated radio and optical networks, which is composed of several modules arranged on the cloud platform and edge nodes. We discuss the function and workflow of each module in detail.

2) An edge-cloud collaboration based traffic prediction mechanism is proposed with the adoption of GRU, which can effectively improve the accuracy of traffic prediction, thereby contributing to the resource allocation scheme.

3) Based on the traffic prediction, an efficient resource allocation method is further proposed. The service reconfiguration and migration process are triggered by the load balancing technology, and the transmission path with the minimum resource consumption is selected to achieve the purpose of resource-saving in integrated radio and optical network scenarios.

The rest of the paper is organized as follows. A summary of the literature is provided in Section II. In Section III, we present the model with edge and cloud architecture considered in this paper and define the relations among the involved nodes and neural networks. In Section IV, we explain how the edge-cloud collaboration traffic prediction takes place. Section V designs the derived resource allocation algorithm. Further, in Section VI, we present the dataset and define the metrics used for the evaluation, and we also evaluate the performances of our algorithm. Finally, Section VII concludes this manuscript.

II. RELATED WORKS

With the improvements in network infrastructure and the high demands of network users, the network scale continues expanding, and multiple service providers and network operators coexist, leading to the emergence of tidal flow effects. Although the tide peaks can be predicted as many as possible, they still may bring uncertainty to network operation and maintenance [15]. In order to solve this problem, a reconfigurable network supporting a software-defined network (SDN) has recently attracted much attention and can adapt to the service demand flow as much as possible [16], [17]. The general reconfiguration framework based on SDN technology consists of two parts, including modeling and forecasting traffic demand flow, and using prediction for active (offline) network optimization between predefined (reconfiguration) time points [18]. The overall goal is to find a resource allocation strategy that is most suitable for the future traffic demand of network [19].

Generally, the previous works mainly have focused on optimizing the efficiency of the resource allocation scheme, without considering the uniformity and sustainability of occupied resources [17], [18], 错误!未找到引用源。, [21]. In this case, although the resource allocation strategy can be solved in a very short time, it is easy to lead to highly unbalanced allocation, and especially after the network reconfiguration, each edge node does not pay enough attention to the quality of service for services.

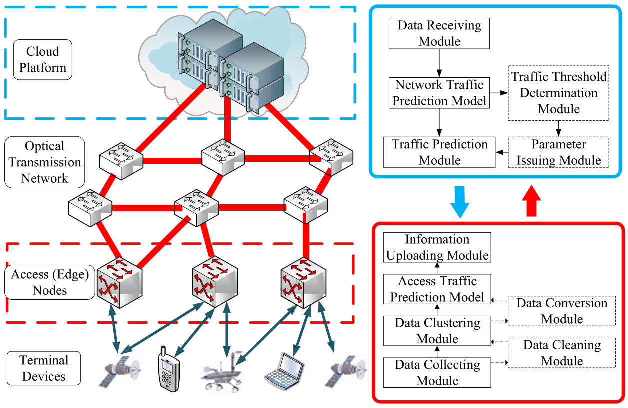

FIGURE 1. The proposed system architecture in integrated radio and optical networks.

In addition, the existing machine learning-assisted traffic prediction is usually based on a unilateral perspective, either cloud or edge, without considering the complementary perspective of the edge and cloud [22], [23]. Prediction based on unilateral perspective may lead to inappropriate resource allocation decisions. Specifically, a unified analysis in cloud via uploading massive raw data not only wastes considerable transmission resources but also fails to provide a timely response for terminals. Nevertheless, the storage ability of edge nodes is insufficient to further improve the accuracy of prediction, which relies heavily on historical data [24].

Both edge and cloud participation in traffic prediction have been considered in [25]. This work analyzed the characteristics of resource demand and load and proposed an adaptive selection strategy and error adjustment factor to select a better prediction algorithm based on a dynamic threshold. Additionally, a short-term forecast of resource demand on the cloud platform was developed. However, the prerequisite is that all access nodes need to regularly report a large amount of traffic data information to the cloud platform for analysis. This step additionally occupies a large amount of transmission pipeline resources, which affects the regular performance of the network. Moreover, the authors of [26] disclosed a traffic forecasting approach for optical backbone network traffic scheduling optimization, which integrates RNN and GRU at the access node for long-term (1 hour in advance) traffic prediction. However, this proposal used several recursive short-term prediction processes (5 minutes each) to obtain the alternative effect of long-term traffic prediction, which inevitably led to cumulative errors and low reliability.

Motivated by the drawbacks above, the traffic prediction of edge cloud collaboration with GRU is explicitly considered in this paper. The purpose is to break the boundary of resource sharing between cloud and edge nodes, so as to provide a more responsive and accurate resource allocation strategy. Although there have been various resource allocation methods using the results of traffic prediction, load balancing algorithms and their related service reconfiguration mechanism have been the focus of network research in recent years. Specifically, load balancing resource allocation scheme has been studied in a wireless network, SDN IP network, and transmission control protocol (TCP) network [13], [15], [16], [27], [28], [29]. However, to the best of our knowledge, few load balancing schemes consider the traffic prediction results based on edge-cloud collaboration in integrated radio and optical networks.

III. SYSTEM MODEL

A. NETWORK ARCHITECTURE

As shown in Fig. 1, the system of integrated radio and optical networks considered in this paper includes a cloud platform, at least one access (edge) node, and at least one terminal device, where an optical transmission network supports the transmission channel.

Cloud Platform refers to an abstraction of the underlying infrastructure of the integrated radio and optical networks, the resources and services of which may be obtained over a transmission network by components of the integrated radio and optical networks through network virtualization techniques.

Access (Edge) Node is a network node that provides mutual access between a wireless workstation and a wired local area network. The cloud platform and the access node are connected through the transmission network. The transmission network provides transparent transmission channels for various services [30]. The switching devices in the transmission network are referred to as transmission network nodes. Their function is to exchange data streams, including distributing and receiving data traffic, over the transmission channels through Ethernet ports. Therefore, an access node is connected to at least one of the transmission network nodes. The cloud platform is also connected to at least one transmission network node in the transmission network.

Terminal Device is a part of a physical layer of integrated radio and optical network architecture. In an IoT system, for example, it may include a temperature and humidity sensor, a QR code tag, a radio frequency identification (RFID) tag, a reader-writer, a camera, a global positioning system (GPS) and other perception terminals. The terminal device may also be connected to at least one access node through a wireless channel [31].

B. STRUCTURE OF EDGE-CLOUD COLLABORATION

As shown in Fig. 1, an access node may include a data collecting module, a data clustering module, an access traffic prediction model, and an information uploading module.

Data Collecting Module is to collect traffic data of the access node through NetFlow technology. The traffic data of the current access node, after being collected, is stored in a database in the format of a tuple, i.e., <date, time, duration, size of traffic, source node IP, destination node IP, source node port, destination node port, collection node ID>.

TABLE 1. Mathematical definitions.

Symbol	Definition
$d_{\mathrm{t}}^{\mathrm{network,p,n}}$	The network traffic of node *p* on port *n* at moment *t*
$r_{\mathrm{t+1}}^{\mathrm{access,p}}$	The prediction result of access traffic of node *p* at moment *t+1*
$r_{\mathrm{t+1}}^{\mathrm{network,p,n}}$	The prediction result of network traffic of node *p* on port *n* at moment *t+1*
$h_{\mathrm{t}}^{\mathrm{p}}$	The traffic threshold set for node *p* corresponding to the moment *t*
$s_{\mathrm{t}}^{\mathrm{p}}$	The size of the network traffic data on node *p* at moment *t*
M_p	The physical traffic limit of node *p*
$k_{\mathrm{t+1}}^{\mathrm{p}}$	The size of the local data required to be uploaded to the cloud platform by node *p* at moment *t+1*

Data Clustering Module is to cluster the traffic data collected by the data collecting module according to the path information. Specifically, it determines whether the traffic data is the access traffic data or the network traffic data based on the source node IP and the destination node IP in the path information of the traffic data. In addition, if traffic data information is generated between the access node and a transmission network node or between access nodes, the path information thereof needs to be recorded. However, if it is generated between a terminal device and the access node, the path information is directly marked as the access traffic. For example, it can be recorded as *Nan*, without recording the IP address and port information of the source/destination node of the traffic, in order to save storage cost of edge nodes.

For easy understanding, we will denote some definitions that may be frequently mentioned in the following as symbols, which are summarized in Table 1.

Access Traffic Prediction Model is to take the access traffic data as input, and output a prediction result $r_{t+1}^{access,p}$ of access traffic at the next moment. It adds the access traffic data to a training set and perform a real-time training to update the access traffic prediction neural network.

Information Uploading Module is to upload the prediction result $r_{t+1}^{access,p}$ output by the access traffic prediction model as well as the network traffic data to the cloud platform. The volume of data to be forwarded will be measured first, by comparing $r_{t+1}^{access,p}$ with the traffic threshold h_t^p. When $r_{t+1}^{access,p} < h_t^p$, the information uploading module will upload $r_{t+1}^{access,p}$ as well as the network traffic data to the cloud platform. When $r_{t+1}^{access,p} \geq h_t^p$, the information uploading module will back up the network traffic data locally at first, then upload $r_{t+1}^{access,p}$ and the size of the network traffic data s_t^p to the cloud platform. It should be noted that the traffic threshold h_t^p is determined and issued to each access node by the cloud platform according to (1) where M_p denotes the physical traffic limit for each access node and $r_{t-1}^{access,p}$ denotes the prediction result of network traffic at the last moment t-1.

$$h_t^p = M_p - r_{t-1}^{access,p} \tag{1}$$

Furthermore, the information uploading module also uploads the network traffic data backed up locally to the cloud platform in chronological order according to the size of the local data required to be uploaded to the cloud platform at the next moment issued from the cloud platform. Such an operation can make the best of the advantage of storage resource of the cloud platform by transferring the storage task of the access node without the heavy burdens of forwarding data to the cloud platform. In this way, limited edge computing resources of the access node can be spared, and the network operation and maintenance costs can be reduced.

The access node may further include a data cleaning module between the data collecting module and the data clustering module, and a data conversion module between the data clustering module and the access traffic prediction model. The former is to clean incomplete, repeated data records and records roaming to the local in the traffic data information collected by the data collecting module. In this solution, the traffic data information is sample-edited before being upload to the cloud platform, so that the burden on transmission channel can be greatly reduced while reserving effective information. The latter is to convert the access traffic data into a data format of a training set for the access traffic prediction model.

As shown in Fig. 1, a cloud platform may include a data-receiving module, a network traffic prediction model, and a traffic prediction module.

The data-receiving module is to receive network traffic data reported by each access node and the prediction result of access traffic at the next moment through the transmission network.

The network traffic prediction model is to take the network traffic reported by each access node as input, and output the prediction result of traffic to be forwarded of node p on port n, which can be denoted as $r_{t+1}^{network,p,n}$. Obviously, we have (2).

$$r_{t+1}^{network,p} = \sum\nolimits_n r_{t+1}^{network,p,n} \tag{2}$$

Further, the network traffic prediction model also regularly adds the network traffic data to the training set and performs real-time training in order to update the network traffic prediction model.

The traffic prediction module is to determine and output the prediction result of traffic at the next moment for each access node, respectively according to the prediction result $r_{t+1}^{access,p}$ reported by each access node, and the prediction result $r_{t+1}^{network,p}$ calculated by cloud and data to be uploaded to the cloud platform at the next moment.

In examples of the present disclosure, the prediction result of traffic at the next moment for the access node p may be determined through the following equation.

$$r_{t+1}^{p} = r_{t+1}^{network,p} + r_{t+1}^{access,p} \tag{3}$$

The cloud platform may further include a traffic threshold determination module and a parameter issuing module. The former is to determine the traffic threshold for each access node corresponding to the next moment according to the prediction result of network traffic for each access node, respectively.

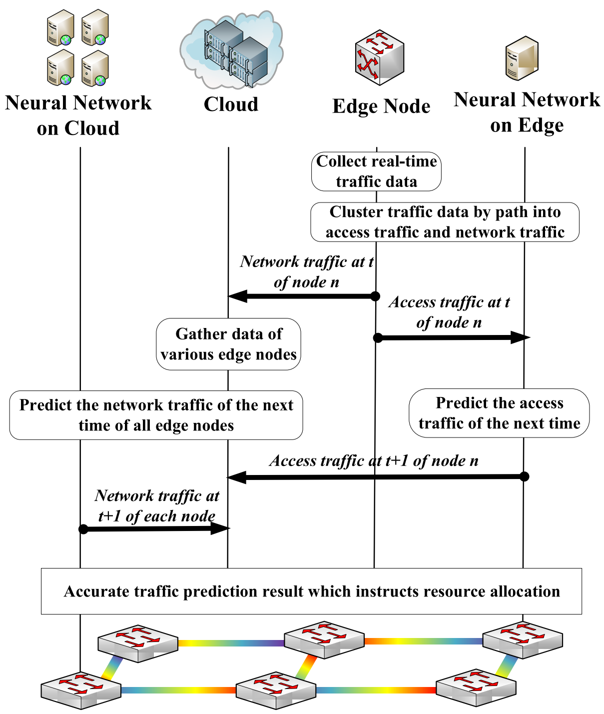

FIGURE 2. Procedure of TP-ECC.

IV. TRAFFIC PREDICTION BASED ON EDGE CLOUD COLLABORATION

A. PRINCIPLE OF NEURAL NETWORKS IN EDGE-CLOUD COLLABORATION

In this subsection, we illustrate that in the integrated radio and optical networks with edge-cloud architecture considered in this paper, how the neural networks are deployed and enabled on cloud and edge nodes, respectively.

The access traffic collected by each edge node is denoted as a time series. For the neural network deployed at each edge node, the primary task is to take the access traffic series of its host node as a training set, and to predict the possible access traffic value at the next moment by time iterations.

Obviously, the prediction accuracy is supported by the amount of data, but it also needs to occupy the resource of excessive storage for historical data. Therefore, we appropriately shorten the time window of these neural networks. In other words, truncate backpropagation through time, and hand over the tasks of storage as well as learning long-term traffic characteristics to cloud nodes.

For the neural network deployed at the cloud platform, it learns the traffic pattern of the entire network topology from a macro perspective. The pre-processed traffic information regularly uploaded by each edge node is used as a training set. The neural network will iteratively acquire multiple other input features besides traffic data, including the connection relationship between nodes and network macro events, and so on. This is a significant improvement to the weakness of the edge nodes with their insufficient computing power and limited vision.

B. TRAFFIC PREDICTION BASED ON EDGE CLOUD COLLABORATION WITH GRU MODEL

We present a traffic prediction strategy based on edge-cloud collaboration (TP-ECC) with GRU model. To capture the feature of traffic, we first construct the GRU model, which can be expressed as follows.

$$u_t = \sigma\left(W_u\left[x_t, h_{t-1}\right] + b_u\right) \tag{4}$$
$$r_t = \sigma\left(W_r\left[x_t, h_{t-1}\right] + b_r\right) \tag{5}$$
$$c_t = \tanh\left(W_c\left[x_t, (r_t \cdot h_{t-1})\right] + b_c\right) \tag{6}$$
$$h_t = u_t \cdot h_{t-1} + (1 - u_t) \cdot c_t \tag{7}$$

where u_t represents the update gate that is used to control the degree of to which the status information at the previous time is brought into the current status, r_t represents the reset gate that is used to control the degree of ignoring the status information at the previous moment, c_t represents the memory content stored at time t, h_{t-1} represents the historical state at time t-1. x_t and h_t represent the input and output state at time t. W and b are the weights and biases in the GRU training process.

Then, the entire procedure is shown in Fig. 2. We will elaborate on the traffic prediction process with the help of mathematical formulas, which can be broken down into the following steps.

Step 1, each access node collects traffic data, respectively.

During the process of traffic prediction, the data collecting module of each edge node detects and records the raw data of traffic flow sequence of access terminals in real time. Each access node may collect time series data as the traffic data through NetFlow technology.

Step 2, each access node clusters the collected traffic data according to the path information thereof and classifies the collected traffic data into the access traffic data and the network traffic data.

Step 3, each access node inputs the access traffic data into the access traffic prediction model configured thereon, respectively, to obtain the prediction result of access traffic at the next moment $r_{t+1}^{access,p}$ output by the traffic prediction model.

Step 4, each access node uploads the prediction result of access traffic at the next moment $r_{t+1}^{\text{access},p}$ and the network traffic data to the cloud platform, respectively.

Firstly, for a port n at an access node, the traffic threshold corresponding to the next moment is calculated according to the prediction result of traffic $r_{t+1}^{network,p,n}$ to be forwarded at the next moment through the following equation.

$$h_{t+1}^{p,n} = m_0 - \sum\nolimits_i r_{t+1}^{network,p,i}, i \in I \tag{8}$$

where m_0 is the physical maximum bearing traffic limit for the access node. I is a set of ports having path dependencies with the port n.

Secondly, the traffic threshold determination module calculates the traffic threshold for the port n of the access node p corresponding to the next moment according to the traffic threshold for the port n of the access node p corresponding to the next moment through the following equation, where N is a set of all ports of the access node p.

$$h_{t+1}^{p} = \sum\nolimits_n h_{t+1}^{p,n}, n \in N \tag{9}$$

Step 5, the cloud platform receives network traffic data and the prediction result of access traffic at the next moment reported by each access node through the transmission network. A macro traffic prediction that considers much more features is going to be performed in the cloud.

Meanwhile, the size of the local data required to be uploaded to the cloud platform at the next moment by each access node is determined by the parameter issuing module through the following method.

When $h_{t+1}^{p} > 0$, the size of the local data required to be uploaded to the cloud platform by the access node at the next moment may be set as (10).

$$k_{t+1}^{p} = s_t^p - s_{t-1}^p \tag{10}$$

When $h_{t+1}^{p} \leq 0$, the size of local data required to be uploaded to the cloud platform by the access node at the next moment may be set as $k_{t+1}^{p} = 0$. In this case, the parameter issuing module may further send alarm information to the access node p and establish a standby link, so as to divert the traffic accessing the node p to other access nodes as much as possible.

Step 6, the cloud platform inputs the network traffic data into the network traffic prediction model to obtain the prediction result of network traffic for each access node at the next moment output by the network traffic prediction model. After that, the cloud platform adds the network traffic data to the training set of the network traffic prediction model and performs real-time training to update the network traffic prediction model.

At the same time, the size of the local data required to be uploaded to the cloud platform at the next moment by each access node and the traffic threshold for each access node corresponding to the next moment are issued to the corresponding access node, respectively.

Step 7, the cloud platform determines and outputs the prediction result of traffic for each access node according to the prediction result of access traffic $r_{t+1}^{access,p}$, the prediction result of network traffic $r_{t+1}^{network,p}$, and the size of the data required to be uploaded to the cloud platform k_{t+1}^{p} at the next moment.

The prediction result of traffic at the next moment for the access node p may be determined through the following equation.

$$r_{t+1}^{p} = r_{t+1}^{network,p} + r_{t+1}^{access,p} + k_{t+1}^{p} \tag{11}$$

As can be seen, TP-ECC solves two important pain points of edge nodes, which are crucial to traffic prediction. First, macro network events like link or node failures can trigger a whole body in flow fluctuation. Second, in a fixed topology, the influence of the connection relationship between nodes cannot be ignored when extracting traffic characteristics. The above problems have to, and can only be considered with the perspective and capabilities of cloud.

In TP-ECC, the method of traffic prediction for edge nodes can predict a relatively long-term network traffic trend by utilizing the cloud platform. Also, the method of traffic prediction for edge nodes can predict a relatively short-term access traffic change by utilizing the access nodes. Further, by combining the relatively long-term network traffic trend and the relatively short-term access traffic change, the one-sidedness and limitation caused by a single position of a prediction module in a network and a single time granularity configuration can be avoided. Thus, the accuracy of traffic prediction for the system can be greatly improved. Furthermore, the access traffic prediction model and the network traffic prediction model have good cycle stability, with no significant change in performance after a plurality of tests and experiments.

V. EFFICIENT RESOURCE ALLOCATION SCHEME

In practical scenarios, the network transmission ability is restricted by its own hardware capacity, and the iteration and expansion of infrastructure require high economic costs. Considering the input-output ratio of hardware, many network operators and service providers are deterred. However, the limitation of infrastructure resources is often the main cause of service congestion. Therefore, how to reasonably allocate resources and provide bandwidth for carrying services under the premise of limited resources is the main goal of this resource allocation algorithm [32].

In this section, we propose an efficient resource allocation strategy that makes full use of the prediction result of TP-ECC. Based on the idea of shortest routing, ERAS introduces constructing auxiliary graph routing, which combines the routing process with real-time resources, and simplifies the virtual network mapping process in integrated radio and optical networks. In order to describe the algorithm conveniently, we first establish a simplified schematic network topology and a set of service request sets as an example,

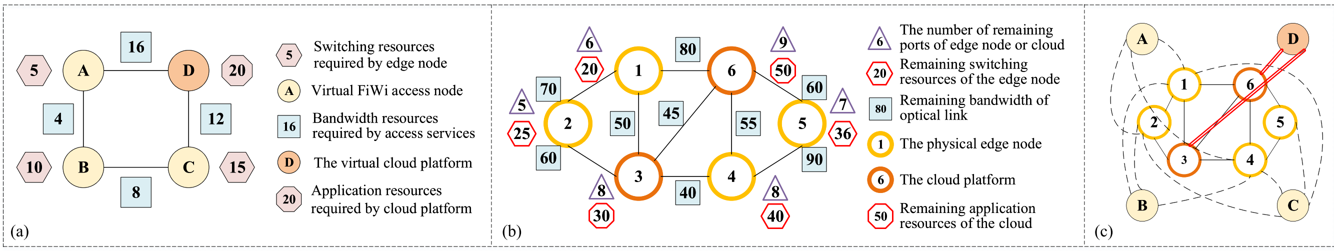

FIGURE 3. **(a) Schematic diagram of business request setting (b) schematic diagram of the optical network of the data center (c) the auxiliary diagram of proposed Efficient Resource Allocation Scheme (ERAS).**

shown in Fig. 3 (a). The yellow circle represents the virtual access node, which is the edge node in edge-cloud collaboration. The orange circle represents the virtual cloud platform, the red hexagon denotes the switching resources required by the virtual edge node, the red octagon is the application resources required by the virtual cloud platform for the perception of network status, and the blue quad indicates the bandwidth resources required by the access service request.

On the basis of the above network topology and service request definition, Fig. 3 (b) shows the schematic diagram of the integrated radio and optical networks with the edge-cloud architecture, in which the yellow circle represents the physical access (edge) node, the orange circle is the cloud platform, the red hollow hexagon represents the remaining switching resources of the physical edge node, the red hollow octagon indicates the remaining application resources of the cloud platform, and the purple triangle represents the number of remaining ports of the physical edge node or cloud platform. The blue quadrilateral represents the remaining bandwidth resource of the optical link.

Fig. 3 (c) shows the auxiliary diagram of ERAS process. First, judge whether the remaining application resources of the cloud platform corresponding to the optimal and suboptimal virtual cloud platform meet the access service requirements in turn. If they meet the requirements, use double solid lines to connect them. If both cloud platforms are lack of resources, the mapping fails, and then traverse all the physical edge nodes. If the remaining switching resources meet the needs of the virtual edge node, use the dotted line to connect them. Otherwise, the link between the node and other nodes is regarded as open circuit.

After drawing the auxiliary diagram, ERAS supplements the weight value for each link. The weight of the dotted line is set to a maximum integer value, which is far more than the total weight of all links in the topology. The weight value of the real line comprehensively considers the demand and residual relationship of nodes and links, as the equilibrium factor Δ, which can be calculated by (12).

$$\Delta = \frac{B_{req}^{i,j}}{B_{res}^{i,j}} + \sqrt{\frac{R_{req}^{i}}{R_{res}^{i}} \cdot \frac{1}{R_{Numres}^{i}+1}}\sqrt{\frac{R_{req}^{j}}{R_{res}^{j}} \cdot \frac{1}{R_{Numres}^{j}+1}} \tag{12}$$

where $B_{req}^{i,j}$ denotes the bandwidth requirement of access services with i and j as source and destination nodes. $B_{res}^{i,j}$ denotes the remaining bandwidth of the optical link with i and j as the source and destination nodes. R_{req}^{i} denotes the number of end-to-end service requests accessed from node i. R_{res}^{i} denotes the remaining bandwidth resources of node i that can be used for forwarding services after TP-ECC prediction.

It can be seen that the request resource is proportional to the weight, and the residual resource is inversely proportional to the weight. This setting follows the idea that the more sufficient the residual resource is, the smaller the weight will be, which is more advantageous in the routing process.

When a service can be mapped to a physical network with multiple routing options, the equilibrium factor of each link is calculated one by one, and the path with the least sum of weights is selected, and the result is associated with the virtual node pair. In the case of severe resource constraints, the priority of traffic distribution to the node with less heavy load to route can maximize the reduction of network blocking rate and traffic delay.

Since the evaluation of cloud platform has been completed in the front stage, the weight value of double solid line is taken as 1 to reduce the impact on the calculation of the shortest path as much as possible. On the basis of the previous step, the auxiliary graph topology after updating the weight is rerouted, and the shortest path is selected for mapping. The whole ERAS algorithm above can be abstractly summarized in Algorithm 1.

In the stage of creating auxiliary graph, the time complexity is $O(m^2)+O(l)$, where m is the number of virtual cloud platforms and l is the number of links in physical topology. In the stage of routing calculation, the shortest path is calculated based on the topology with updated weight, and the time complexity is $O(n^3)$, so the comprehensive time complexity of the algorithm is $O(m^2)+O(l)+O(n^3)$.

VI. PERFORMANCE EVALUATION

A. EXPERIMENTAL SETUP

Experiments in this section are designed to demonstrate the traffic prediction accuracy of TP-ECC as well as the overall performance of ERAS.

Python is used to generate the underlying physical topology [34], in which 200 nodes are evenly distributed in four

Algorithm 1 Efficient Resource Allocation Scheme

Input: Virtualization business request $R_{P_i}(S, D, d, C_p)$;
Output: Virtualization resource allocation mapping results, Success / False;

for each working cycle
  Initialize virtualization mapping results $\boldsymbol{R^{ov}}$ as False.
  for virtual cloud platform d^v in set D **do**
    if the remaining application resources of the corresponding cloud platform meet the demand, that is, $A_r\left(d_i^p\right) > A(d^v)$ **then**
      Use double solid lines to connect the corresponding cloud platform.
    end if
  end for
  for virtual edge node d^v in set D **do**
    if the remaining exchange resources of corresponding nodes meet the demand, that is, $C_r\left(v_i^p\right) > C(v^v)$ **then**
      Use dotted lines to connect the corresponding nodes.
    else
      Remove the physical edge node.
    end if
  end for
  Add weight value for all links.
  The shortest path of virtual network mapping is calculated on the basis of $G^p = \left\{\boldsymbol{P_{nj}}, \boldsymbol{x^*_{nk}}\right\}$ after the link weight is supplemented.
  Set virtualization mapping results $\boldsymbol{R^{ov}}$ as True.
end for
return $\left\{\boldsymbol{P_{nj}}, \boldsymbol{x^*_{nk}}\right\}, \boldsymbol{R^{ov}}$

domains, each domain contains a cloud platform, and the total content in the cloud platform is set to 800. The other switching nodes are divided into advanced switching nodes and middle and low-end switching nodes. In the simulation, the difference between them is mainly reflected in the size of switching capacity. The former's switching capacity is set to 1800, and the latter's switching capacity is set to 400. In addition to the above parameters, the number of ports of the former is 128, and the latter is 10. The resources of the links connecting the above nodes are represented by the free spectrum slots, and the total amount of link resources is set to 358 [35]. We generated traffic data with 4510 optical nodes in the State Key Laboratory in July 2019, comprising over 27,180,000 traffic flows and 3,000 queries around 3.5GB. The services in the network are composed of end-to-end services and networking services. The number of switching resources and application resources requested by nodes are randomly generated, and the link resources requested by virtual links are randomly selected, and the content of service request is randomly selected, the number of service nodes is set as a random integer between 5 and 10. The length of service queue in a single simulation is set to 1000, and the average value of multiple simulation data is taken as the simulation result.

B. ACCURACY OF TRAFFIC PREDICTION

Figure 4(a) shows the prediction results of TP-ECC and the ground traffic series. The API of neural networks used in the TP-ECC algorithm are based on TensorFlow 1.2.1 in Python 2.6. It can be seen that the prediction is generally accurate, and some special phenomena can be further explained. First, the prediction results are not disturbed by any instantaneous traffic reductions, so the future results are still reasonable. This is because the cloud's grasp of the overall trend prevents edge nodes from being over-sensitive to those abnormal reduction data. Second, early predictions are made precisely for the accidental traffic spikes. The reason is that when the storage task is transferred to the cloud, the edge node has more resource to ensure the local prediction is refined and efficient.

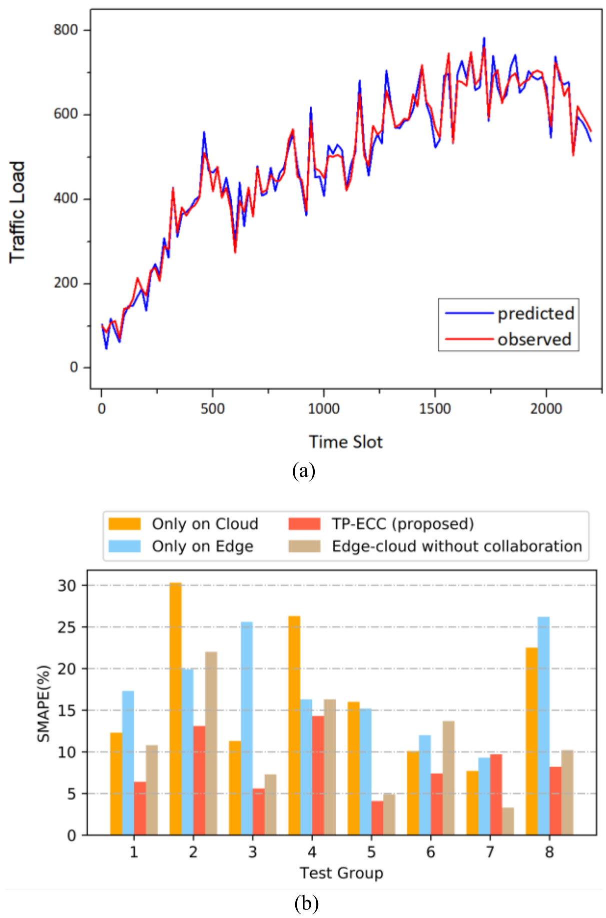

FIGURE 4. Results on traffic prediction: (a) prediction of TP-ECC (b) Comparison on SMAPE.

In Fig. 4(b), by analyzing the Symmetric Mean Absolute Percentage Error (SMAPE) of prediction, we compare the

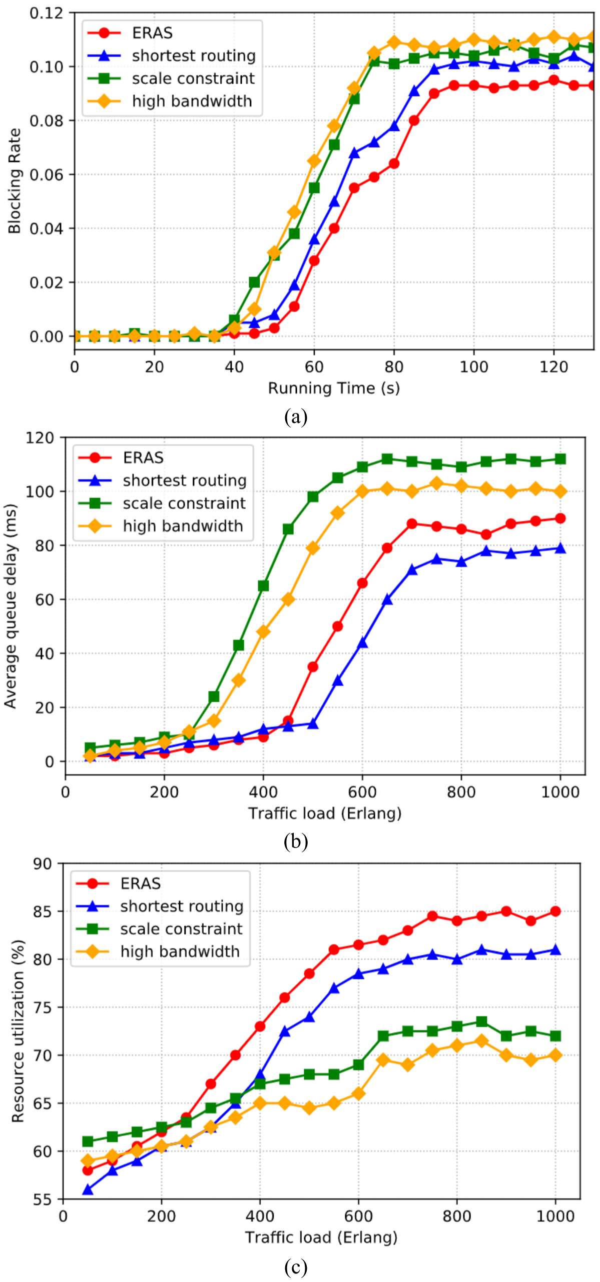

FIGURE 5. Network performance: (a) blocking rate (b) average queue delay (c) resource utilization.

performance of TP- ECC with the methods in which neural networks are deployed in different locations, including on the cloud or edge alone, or both on the cloud and edge but not coordinated. Obviously, the SMAPE of TP-ECC is the lowest in all test groups, which indicates that it has the best accuracy. Quantitatively, it is 8.1%, 9.5%, and 8.8% higher than the average accuracy of the other three methods. This numerically proves that the complementary characteristics of edge nodes and cloud can improve each other by TP-ECC, which is helpful for optical network to establish countermeasures in the traffic fluctuation introduced by large scale access terminal sets in integrated radio and optical networks.

C. NETWORK PERFORMANCE OF ERAS SCHEME

In this section, we analyze the overall performance of ERAS by comparing it with the other three conventional resource allocation algorithms which are the shortest routing algorithm, the scale constraint algorithm and the high bandwidth virtualization algorithm. The evolution is conducted from three aspects including traffic blocking rate, average queue delay and resource utilization.

Figure 5(a) shows the relationship between the traffic blocking rate and the running time under the action of the four algorithms, respectively. It can be seen that all of the four algorithms have experienced a very low blocking rate in the initial stage, and then the blocking rate surges with the passage of time, and finally the increasing rate gradually decreases to a stable level, but at this stage, different algorithms have different blocking rate performance. This is because at the beginning, the network resources are sufficient, and the controller has more space to select the required resources for services. When more services flow in, some resources are occupied but not released, such as the current hot application resources or key switching nodes, resulting in the failure of some subsequent services mapping. In the later period, the occupation and release of network resources tend to be stable, and the blocking rate of services also presents a corresponding trend. In the stable stage, for the end-to-end traffic, the load balancing virtualization algorithm has a lower blocking rate, which is 6.91% lower than the shortest routing algorithm. The analysis shows that it can carry more traffic at the same time because of the optimal path selected after fully considering the resource occupation when calculating the routing.

Then, the simulation data of average queue delay shown in Fig. 5(b) is analyzed. As a whole, no matter which algorithm is used, the average queue delay increases gradually with the increase of traffic intensity. This is because when the resources become tight, the algorithm can only select the optimal path that meets the conditions as far as possible under the current resource occupation. However, in terms of stability, for end-to-end services, the shortest route virtualization algorithm selects the shortest path, which has more advantages in the calculation of routing stage and information transmission nodes, and the average delay is reduced by 15.61% compared with the ERAS algorithm which increases with the increase of traffic intensity.

Finally, we observe the statistics of resource utilization in the network. As shown in Fig. 5(c), with the increase of service intensity, it has experienced a process of first increasing and then gradually stabilizing. In the stable stage, the load balancing virtualization algorithm presents a higher resource utilization, which is 8.64% higher than the shortest routing algorithm. To summarize, for end-to-end access service, ERAS algorithm achieves lower traffic blocking rate and higher resource utilization at the cost of delay.

VII. CONCLUSION

In this paper, we have first proposed a new deployment architecture in order to achieve edge cloud collaboration in integrated radio and optical networks. Then, using the functional entities of the architecture, we have proposed the accurate traffic prediction algorithm TP-ECC which is based on edge cloud collaboration with GRU model. We further have proposed a resource allocation scheme ERAS based on load balancing theory. Their performance has been demonstrated by experiments in integrated radio and optical networks testbed.

We also have evaluated the performance of the proposed algorithm for end-to-end services under heavy traffic load, and compared it with other conventional resource allocation schemes. Numerical results have shown that with large terminal sets, TP-ECC is fully capable to improve the accuracy rates of traffic prediction by up to 9.5%, compared with the methods without edge-cloud collaboration. Furthermore, ERAS can improve the resource utilization of the whole network, while reducing the average queue delay and blocking probability.

In the future, we would like to study more complex resource allocation optimization technology and edge cloud collaboration architecture, which can jointly mobilize the resources of cloud and edge nodes and further enhance the security, reliability, and accuracy of the fast-growing integrated radio and optical networks.

REFERENCES

[1] J. Yu, X. Liu, Y. Gao, and X. Shen, "3D channel tracking for UAV-satellite communications in space-air-ground integrated networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2810–2823, Dec. 2020.

[2] X. Chen, Y. Bi, X. Chen, H. Zhao, N. Cheng, F. Li, and W. Cheng, "Dynamic service migration and request routing for microservice in multicell mobile-edge computing," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13126–13143, Aug. 2022.

[3] H. Yang, X. Zhao, Q. Yao, A. Yu, J. Zhang, and Y. Ji, "Accurate fault location using deep neural evolution network in cloud data center interconnection," *IEEE Trans. Cloud Comput.*, vol. 10, no. 2, pp. 1402–1412, Apr. 2022.

[4] M. Zhao, J. Yu, W. Li, D. Liu, S. Yao, W. Feng, C. She, and T. Quek, "Energy-aware task offloading and resource allocation for time-sensitive services in mobile edge computing systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10925–10940, Oct. 2021.

[5] H. Yang, J. Yuan, C. Li, G. Zhao, Z. Sun, Q. Yao, B. Bao, A. V. Vasilakos, and J. Zhang, "BrainIoT: Brain-like productive services provisioning with federated learning in industrial IoT," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2014–2024, Feb. 2022.

[6] C. Ding, A. Zhou, X. Ma, N. Zhang, C.-H. Hsu, and S. Wang, "Towards diversified IoT services in mobile edge computing," *IEEE Trans. Cloud Comput.*, early access, Sep. 3, 2021, doi: 10.1109/TCC.2021.3109385.

[7] A. Naouri, H. Wu, N. A. Nouri, S. Dhelim, and H. Ning, "A novel framework for mobile-edge computing by optimizing task offloading," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 13065–13076, Aug. 2021.

[8] C. Li, H. Yang, Q. Yao, Z. Sun, and J. Zhang, "High-precision edge-cloud collaboration with federated learning in edge optical network," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2021, pp. 1–3.

[9] H. Ke, H. Wang, W. Sun, and H. Sun, "Adaptive computation offloading policy for multi-access edge computing in heterogeneous wireless networks," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 1, pp. 289–305, Mar. 2022.

[10] D. P. Isravel, S. Silas, and E. B. Rajsingh, "Centrality based congestion detection using reinforcement learning approach for traffic engineering in hybrid SDN," *J. Netw. Syst. Manage.*, vol. 30, no. 1, pp. 1–22, Jan. 2022.

[11] H. Yang, K. Zhan, B. Bao, Q. Yao, J. Zhang, and M. Cheriet, "Automatic guarantee scheme for intent-driven network slicing and reconfiguration," *J. Netw. Comput. Appl.*, vol. 190, Sep. 2021, Art. no. 103163.

[12] R.-K. Shiu, Y.-W. Chen, P.-C. Peng, J. Chiu, Q. Zhou, T.-L. Chang, S. Shen, J.-W. Li, and G.-K. Chang, "Performance enhancement of optical comb based microwave photonic filter by machine learning technique," *J. Lightw. Technol.*, vol. 38, no. 19, pp. 5302–5310, Oct. 1, 2020.

[13] X. Fu and C. Zhou, "Predicted affinity based virtual machine placement in cloud computing environments," *IEEE Trans. Cloud Comput.*, vol. 8, no. 1, pp. 246–255, Jan. 2020.

[14] B. Bao, H. Yang, Y. Wan, Q. Yao, A. Yu, J. Zhang, B. C. Chatterjee, and E. Oki, "Node-oriented traffic prediction and scheduling based on graph convolutional network in metro optical networks," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2021, pp. 1–3.

[15] G. O. Perez, A. Ebrahimzadeh, M. Maier, J. A. Hernandez, D. L. Lopez, and M. F. Veiga, "Decentralized coordination of converged tactile internet and MEC services in H-CRAN fiber wireless networks," *J. Lightw. Technol.*, vol. 38, no. 18, pp. 4935–4947, Sep. 15, 2020.

[16] Z. Ni, H. Chen, Z. Li, X. Wang, N. Yan, W. Liu, and F. Xia, "MSCET: A multi-scenario offloading schedule for biomedical data processing and analysis in cloud-edge-terminal collaborative vehicular networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Nov. 30, 2021, doi: 10.1109/TCBB.2021.3131177.

[17] S. Yin, Y. Chu, C. Yang, Z. Zhang, and S. Huang, "Load-adaptive energy-saving strategy based on matching game in edge-enhanced metro FiWi," *Opt. Fiber Technol.*, vol. 68, Jan. 2022, Art. no. 102762.

[18] E. Ganesan, I.-S. Hwang, A. T. Liem, and M. S. Ab-Rahman, "SDN-enabled FiWi-IoT smart environment network traffic classification using supervised ML models," *Photonics*, vol. 8, no. 6, p. 201, Jun. 2021.

[19] T. Wu, P. Zhou, B. Wang, A. Li, X. Tang, Z. Xu, K. Chen, and X. Ding, "Joint traffic control and multi-channel reassignment for core backbone network in SDN-IoT: A multi-agent deep reinforcement learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 1, pp. 231–245, Jan. 2021.

[20] C. Li, H. Yang, Z. Sun, Q. Yao, B. Bao, J. Zhang, and A. V. Vasilakos, "Federated hierarchical trust-based interaction scheme for cross-domain industrial IoT," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 447–457, Jan. 2023.

[21] Y. Li, X. Sun, H. Zhang, Z. Li, L. Qin, C. Sun, and Z. Ji, "Cellular traffic prediction via a deep multi-reservoir regression learning network for multi-access edge computing," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 13–19, Oct. 2021.

[22] J. Bi, S. Li, H. Yuan, and M. Zhou, "Integrated deep learning method for workload and resource prediction in cloud systems," *Neurocomputing*, vol. 424, pp. 35–48, Feb. 2021.

[23] X. Song, Y. Guo, N. Li, and L. Zhang, "Online traffic flow prediction for edge computing-enhanced autonomous and connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2101–2111, Mar. 2021.

[24] H. Yang, A. Yu, J. Zhang, J. Nan, B. Bao, Q. Yao, and M. Cheriet, "Data-driven network slicing from core to RAN for 5G broadcasting services," *IEEE Trans. Broadcast.*, vol. 67, no. 1, pp. 23–32, Mar. 2021.

[25] J. Chen and Y. Wang, "An adaptive short-term prediction algorithm for resource demands in cloud computing," *IEEE Access*, vol. 8, pp. 53915–53930, 2020.

[26] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4927–4943, Jun. 2022, doi: 10.1109/TITS.2021.3054840.

[27] M. Chen, S. Huang, X. Fu, X. Liu, and J. He, "Statistical model checking-based evaluation and optimization for cloud workflow resource allocation," *IEEE Trans. Cloud Comput.*, vol. 8, no. 2, pp. 443–458, Apr. 2020.

[28] W. Wei, H. Gu, K. Wang, J. Li, X. Zhang, and N. Wang, "Multi-dimensional resource allocation in distributed data centers using deep reinforcement learning," *IEEE Trans. Netw. Service Manage.*, early access, Oct. 11, 2022, doi: 10.1109/TNSM.2022.3213575.

[29] R. Kumar and V. Tiwari, "Opt-ACM: An optimized load balancing based admission control mechanism for software defined hybrid wireless based IoT (SDHW-IoT) network," *Comput. Netw.*, vol. 188, Apr. 2021, Art. no. 107888.

[30] Q. Yao, H. Yang, B. Bao, A. Yu, J. Zhang, and M. Cheriet, "Core and spectrum allocation based on association rules mining in spectrally and spatially elastic optical networks," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5299–5311, Aug. 2021.

[31] H. Yang, B. Bao, C. Li, Q. Yao, A. Yu, J. Zhang, and Y. Ji, "Blockchain-enabled tripartite anonymous identification trusted service provisioning in industrial IoT," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2419–2431, Feb. 2022.

[32] B. Bao, H. Yang, Q. Yao, A. Yu, B. C. Chatterjee, E. Oki, and J. Zhang, "SDFA: A service-driven fragmentation-aware resource allocation in elastic optical networks," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 1, pp. 353–365, Mar. 2022.

[33] X. Long, J. Wu, and L. Chen, "Energy-efficient offloading in mobile edge computing with edge-cloud collaboration," in *Proc. Int. Conf. Algorithms Archit. Parallel Process.* Cham, Switzerland: Springer, 2018, pp. 460–475.

[34] Z. Sun, H. Yang, C. Li, Q. Yao, D. Wang, J. Zhang, and A. V. Vasilakos, "Cloud-edge collaboration in industrial Internet of Things: A joint offloading scheme based on resource prediction," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17014–17025, Sep. 2022.

[35] H. Yang, Q. Yao, B. Bao, A. Yu, J. Zhang, and A. V. Vasilakos, "Multi-associated parameters aggregation-based routing and resources allocation in multi-core elastic optical networks," *IEEE/ACM Trans. Netw.*, vol. 30, no. 5, pp. 2145–2157, Oct. 2022.

BOWEN BAO received the M.S. degree in computer science and technology from the Hebei University of Engineering, Handan, China, in 2019. He is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests include elastic optical networks, spectrum assignment and routing, fragmentation, distance-adaptive transmission, and physical layer impairments.

HUI YANG (Senior Member, IEEE) is currently the Vice Dean and a Professor at the Beijing University of Posts and Telecommunications (BUPT). His research interests include SDN, AI, elastic optical networks, and fixed-mobile access networks. He has authored or coauthored 100 papers in prestigious journals and conferences, and he is the first author of more than 50 of them. He received the Best Paper Award at IWCMC' 19/NCCA' 15 and the Young Scientist Award at IEEE ICOCN' 17. He was the General Chair of ISAI' 16. He has served as the Guest Editor for IEEE Journal on Selected Areas in Communications and an Associate Technical Editor for *IEEE Communications Magazine*.

QIUYAN YAO (Member, IEEE) received the M.S. degree in computer science and technology from the Hebei University of Engineering, Handan, China, in 2015, and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2020. She is currently working at BUPT. Her research interests include the AI driven routing, spectrum assignment strategy in elastic optical networks, and space division multiplexing networks. She received the Best Paper Award at OECC/PSC' 19.

LIN GUAN received the M.S. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2021. Her research interests include the artificial intelligence model, edge-cloud collaboration, and resource allocation.

JIE ZHANG (Member, IEEE) is currently a Professor and the Dean of the Institute of Information Photonics and Optical Communications, BUPT. He is sponsored by more than ten projects of the Chinese government. He has published eight books and more than 100 articles. Seven patents have also been granted. He has served as a TPC Member for ACP 2009, PS 2009, and ONDM 2010. His research interests include optical transport networks and packet transport networks.

MOHAMED CHERIET (Senior Member, IEEE) is currently a Full Professor with the Department of System Engineering, Ecole de Technologie Supérieure, Montreal, QC, Canada. He has authored or coauthored more than 300 technical papers in renowned international journals and conferences, and has delivered more than 50 invited talks. He was a recipient of the 2016 IEEE Canada J. M. Ham Outstanding Engineering Educator Award, the 2013 ETS Research Excellence Prize, and the 2012 Queen Elizabeth Diamond Jubilee Medal. He is a fellow of IAPR, CAE, EIC, and EC.